An Information-Theoretic Analysis of Thompson Sampling

Daniel Russo, Benjamin Van Roy

University of Texas at Austin

Presented by Kelsey Ball, Joshua Ong, Harshit Sikchi, Khang Le, Ruichen Jiang

September 24, 2022

Outline

1. An introduction to bandits
2. Thompson Sampling
3. General Regret Bound
4. Information Ratio Bounds
5. Extension to bandits with many actions

Problem Setting

- Player has a set of actions, each with unknown stochastic reward
- In each round, the player chooses an action and observes some reward
- Player wants to maximize reward over time

Why "Multi-Armed Bandits"?

Figure: A one-armed bandit.

Why “Multi-Armed Bandits”?

Figure: A multi-armed bandit.

A Simple Game

$\boxed{\text{Coin 1}}$

$\Pr(H) = ??$

$\boxed{\text{Coin 2}}$

$\Pr(H) = ??$

In each round $t = 1, \ldots, T$:

1. Choose a coin
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0.

Goal: maximize winnings.

A Simple Game

$\boxed{\text{Coin 1}}$

$\Pr(\mathrm{H}) = \theta_1 = ??$

$\boxed{\text{Coin 2}}$

$\Pr(\mathrm{H}) = \theta_2 = ??$

A simple strategy:

1. Flip each coin 10 times, then estimate bias using sample mean:

$$\hat{\theta}_i = \frac{\#\ \text{Heads, coin i}}{10}$$

2. For remaining rounds, if $\hat{\theta}_1 \geq \hat{\theta}_2$: play coin 1; else play coin 2

A Simple Game

$\boxed{\text{Coin 1}}$

$\Pr(\mathrm{H}) = \theta_1 = ??$

$\boxed{\text{Coin 2}}$

$\Pr(\mathrm{H}) = \theta_2 = ??$

A simple strategy:

1. $\underbrace{\text{Flip each coin 10 times}}_{\text{Exploration}}$, then estimate bias using sample mean
2. For remaining rounds, $\underbrace{\text{if } \hat{\theta}_1 \geq \hat{\theta}_2\text{: play coin 1; else play coin 2}}_{\text{Exploitation}}$

A Simple Game

$\boxed{\text{Coin 1}}$

$\Pr(\mathrm{H}) = ??$

$\boxed{\text{Coin 2}}$

$\Pr(\mathrm{H}) = ??$

In each round $t = 1, \ldots, T$:

1. Choose a coin
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0.

Goal: maximize winnings.

Some Terminology

Coin 1	Coin 2
$\mathrm{Bern}(\theta_1)$	$\mathrm{Bern}(\theta_2)$

"Environment" $\theta = [\theta_1, \theta_2]$

In each round $t = 1, \ldots, T$:

1. Choose a coin
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0.

Goal: maximize winnings.

Some Terminology

$\boxed{\text{Coin 1}}$ $\boxed{\text{Coin 2}}$

$\mathrm{Bern}(\theta_1)$ $\mathrm{Bern}(\theta_2)$

"Environment" $\theta = [\theta_1, \theta_2]$

In each round $t = 1, \ldots, T$:

1. Choose a coin $\leftarrow$ "Action" or "Arm" A_t
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0.

Goal: maximize winnings.

Some Terminology

Coin 1 | Coin 2

$\mathrm{Bern}(\theta_1)$ | $\mathrm{Bern}(\theta_2)$

"Environment" $\theta = [\theta_1, \theta_2]$

In each round $t = 1, \ldots, T$:

1. Choose a coin $\leftarrow$ "Action" or "Arm" A_t
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0. $\leftarrow$ "Reward" $R(A_t, \theta)$

Goal: maximize winnings.

Some Terminology

$\boxed{\text{Coin 1}}$ $\boxed{\text{Coin 2}}$

$\mathrm{Bern}(\theta_1)$ $\mathrm{Bern}(\theta_2)$

"Environment" $\theta = [\theta_1, \theta_2]$

In each round $t = 1, \ldots, T$:

1. Choose a coin $\leftarrow$ "Action" or "Arm" A_t
2. Observe heads or tails
3. If heads, win \$1. If tails, win \$0. $\leftarrow$ "Reward" $R(A_t, \theta)$

Goal: maximize winnings. $\leftarrow$ Equivalently, minimize "Regret"

What Is Regret?

$\boxed{\text{Coin 1}}$ $\qquad$ $\boxed{\text{Coin 2}}$

$\mathrm{Bern}(\theta_1)$ $\qquad$ $\mathrm{Bern}(\theta_2)$

$$\text{Environment } \theta = [\theta_1, \theta_2]$$

In each round $t = 1, \ldots, T$:

1. Choose A_t
2. Observe $R(A_t, \theta)$

Note that the optimal action is $A^* = \underset{i}{\text{argmax}}\ \theta_i$.

Regret is the following reward gap:

$$\text{Regret} = R(A^*, \theta) - R(A_t, \theta)$$

Defining Regret

$$\text{Instantaneous Regret} = R(A^*, \theta) - R(A_t, \theta)$$

$$\text{Total Regret} = \sum_{t=1}^{T} R(A^*, \theta) - R(A_t, \theta)$$

$$\text{Total Expected Regret} = \underbrace{E\left[\sum_{t=1}^{T} R(A^*, \theta)\right]}_{\substack{\text{expected reward of} \\ \text{optimal policy}}} - \underbrace{E\left[\sum_{t=1}^{T} R(A_t, \theta)\right]}_{\substack{\text{expected reward of} \\ \text{player's policy}}}$$

Bayesian Bandits

Now suppose there are two possible sets of coins:

Coin 1	Coin 2
Bern(0.6)	Bern(0.5)

Environment 1

Coin 1	Coin 2
Bern(0.7)	Bern(0.2)

Environment 2

Bayesian Bandits assumes a **prior distribution P over environments:**

$$P = \begin{cases} \text{Environment 1 w.p. } \frac{3}{4} \\ \text{Environment 2 w.p. } \frac{1}{4} \end{cases}$$

At time t=0, a ground-truth environment $\theta \sim P$ is sampled, which generates the rewards from t=1 onward.

The player knows P, but does not know θ.

Frequentist vs. Bayesian Regret

$$\text{Frequentist Regret} = E\left[\sum_{t=1}^{T} R(A^*, \theta)\right] - E\left[\sum_{t=1}^{T} R(A_t, \theta)\right]$$

$$= \text{Reg}_T(\underbrace{\pi}_{\text{policy}}, \underbrace{\theta}_{\substack{\text{fixed}\\ \text{environment}}})$$

$$\text{Bayesian Regret} = \underset{\theta \sim P}{E}\left[\text{Reg}_T(\pi, \theta)\right]$$

where P is the distribution over possible environments.

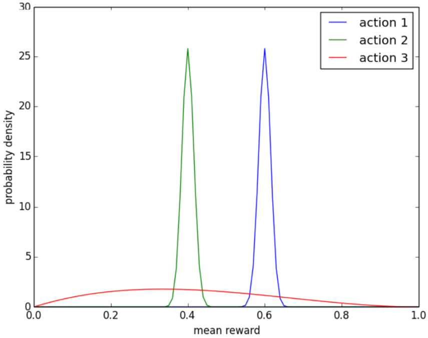
This talk will focus on **Bayesian** regret.

Mini bandit Q&A

Another Motivating Example

Bayesian Setting

1. $Bern(\theta_k) = [.6, .4, ?]$
2. Sampled arms $n_i = [1000, 1000, 10]$
3. prior belief $p_k \sim$ Beta-distribution$(\alpha_k, \beta_k)(\theta_k)$ that model uncertainty.



[1]Russo, D.; Van Roy, B.; Kazerouni, A.; Osband, I. Wen, Z. A Tutorial on Thompson Sampling arXiv, 2017

A Greedy Algorithm

Question?

What is a greedy way to minimize regret?

Greedy Algorithm

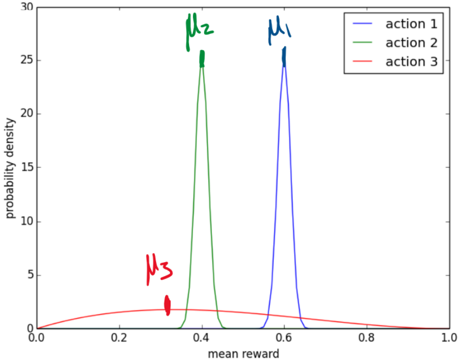
1. for $t \in T, k \in K$
2. $\hat{\mu_k} \leftarrow \mathbb{E}_{p_k}[\theta]$ where $p_k \sim Beta_{\alpha,\beta}(\theta)$ // take means
3. $a_t \leftarrow \text{argmax}_k \hat{\mu_k}$ // greedy step
4. play action a_t and observe reward $x_t = R(a_t, \theta^*)$
5. $p \leftarrow \mathbb{P}(\theta | a_t, x_t) \approx Beta_{\alpha+x_t, \beta+1-x_t}(\theta)$ //update posterior

Note: the nice thing about the Beta distribution is that is has a closed form posterior update that is another Beta distribution, in general this is not the case. $\mathbb{E}_{p_k}[\theta] = \alpha_k/(\alpha_k + \beta_k)$.

Greedy Continued.

Greedy Perspective

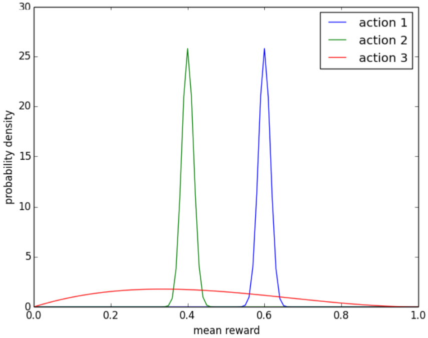
1. Arms give rewards $\sim Bern(\theta_i)$
2. θ_i is parameterized by Beta-distribution.
3. Greedy takes the expectation of the priors: $Bern(\hat{\mu}_1) > Bern(\hat{\mu}_3)$



Motivating Thompson Sampling

Algorithms

1. Greedy
2. $\epsilon-$Greedy
3. Thompson Sampling



Thompson Sampling Algorithm

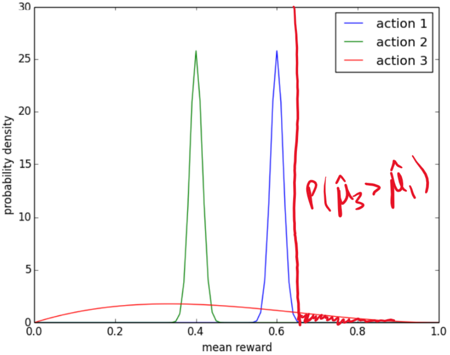
~~Greedy Algorithm~~ TS Sampling

1. for $t \in T, k \in K$
2. $\cancel{\hat{\mu}_k \leftarrow \mathbb{E}_{p_k}[\theta]}$ ~~where~~ $\cancel{p_k \sim Beta_{\alpha,\beta}(\theta)}$
   $\hat{\mu}_k \sim p_k$ where $p_k \sim Beta_{\alpha,\beta}(\theta)$ //sample posterior
3. $a_t \leftarrow \text{argmax}_k \hat{\mu}_k$
4. play action a_t and observe reward $x_t = R(a_t, \theta^*)$
5. $p \leftarrow \mathbb{P}(\theta | a_t, x_t) \approx Beta_{\alpha+x_t,\beta+1-x_t}(\theta)$ //update posterior

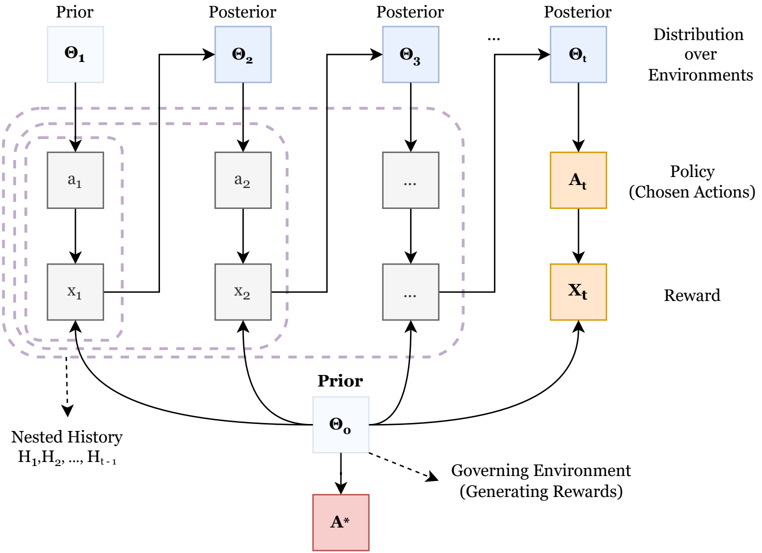
Thompson Sampling Continued.

Thompson Perspective

1. $P(\ \hat{\mu}_1 > \hat{\mu}_2)$?
2. $P(\ \hat{\mu}_1 > \hat{\mu}_3)$?
3. balancing exploration vs. exploitation



Visual Representation of Thompson Sampling



Some more technical details. (beyond Bernoulli)

Notation

1. Approximate posterior updates for when updates are intractable
   1. Gibbs sampling, sampling from a Laplace approximation and more. [1]
2. Bayesian vs. Frequentist implementation of TS [2]

---

[1] Russo, D.; Van Roy, B.; Kazerouni, A.; Osband, I. Wen, Z. A Tutorial on Thompson Sampling arXiv, 2017

[2] Lattimore, T. Szepesvifmmodeaelseáfiri, C. Bandit Algorithms Cambridge Core, Cambridge University Press, 2020

Hook: Analysis with Information Theory

ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

BY WILLIAM R. THOMPSON. From the Department of Pathology, Yale University.

Section 1.

IN elaborating the relations of the present communication interest was not centred upon the interpretation of particular data, but grew out of a general interest in problems of research planning. From this point of view there can be no

Invented in 1933.

$\longrightarrow$

Tweet

Maximilian Kasy
@maxkasy

Reading rec (technical):
"An information-theoretic analysis of Thompson sampling" by Russo & Van Roy.
dl.acm.org/doi/abs/10.555...
A beautiful derivation of performance guarantees for adaptive decision algorithms, relating welfare (regret) and information acquisition.

ıen an action is sampled, a random reward in $[0,1]$ is re
iven by (1). Then, our analysis establishes that the exp
on sampling up to time T is bounded by

$$\sqrt{\frac{\text{Entropy}(A^*)dT}{2}},$$

Analyzed in 2016.

Mini Thompson Sampling Q&A

Our Goal in Online Learning

Notations

1. A_t: action played by the algorithm at time t.
2. A^*: optimal action that could have been played.
3. θ: environment parameters
4. $R(A_t, \theta)$: reward you get from playing action A_t in environment θ.
5. $\mathcal{H}_t$: History of action and rewards seen till time t. $(A_1, R_1, A_2, R_2, .., A_t, R_t)$
6. h_t: A particular history of action and rewards till time t. $(a_1, r_1, a_2, r_2, ..., a_t, r_t)$

Note that all of the above are random variables. For example A^* becomes completely known if we know the environment θ.

Our Goal in Online Learning

Goal: Minimize Bayesian Regret till time T

$$\min \mathbb{E}_{\theta \sim P_\theta} \left[\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_{t-1}|\theta}[R(A^*, \theta) - R(A_t, \theta)] \right] \tag{1}$$

The inner expectation is taken over possible histories $\mathcal{H}_{t-1}$ produced by our online learner and the environment.

Our Goal in Online Learning

Equivalent definitions of Bayesian Regret

Let θ, h_{t-1} come from a joint distribution denoted by $P_{\theta,\mathcal{H}_{t-1}}$. The marginals are denoted by P_θ and $P_{\mathcal{H}_{t-1}}$

$$\mathbb{E}_{\theta\sim P_\theta}\left[\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_{t-1}|\theta}[R(A^*,\theta)-R(A_t,\theta)]\right] \quad \text{(Original Definition)} \tag{2}$$

$$=\sum_{t=1}^{T}\mathbb{E}_{(\theta,h_{t-1})\sim P_{\theta,\mathcal{H}_{t-1}}}[R(A^*,\theta)-R(A_t,\theta)] \tag{3}$$

$$=\sum_{t=1}^{T}\mathbb{E}_{h_{t-1}\sim P_{\mathcal{H}_{t-1}}}\left[\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta)-R(A_t,\theta)]\right] \tag{4}$$

We will be most interested in the last definition of Bayesian Regret.

Information Theory Basics

Mutual Information Definition

$$I(X;Y) = H(X) - H(X|Y) \tag{5}$$

Chain Rule for Mutual Information

$$I(X;(Z_1, Z_2..., Z_T)) = I(X;Z_1) + I(X;Z_2|Z_1) + .. + I(X;Z_T|Z_1, .., Z_{T-1}) \tag{6}$$

Conditional Mutual Information to Mutual Information

$$\mathbb{E}_{z \sim Z}[I(X|Z=z)] = I(X|Z) \tag{7}$$

Information Ratio - Definition

Information Ratio

$$\Gamma_t = \frac{\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta) - R(A_t,\theta)]^2}{I(A^*;(A_t, R(A_t,\theta))|h_{t-1})} \tag{8}$$

Numerator: $\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta) - R(A_t,\theta)]$

The difference between the the best reward I can get vs the reward I actually get given I play based on history knowledge h_{t-1}.

Denominator: $I(A^*;(A_t, R(A_t,\theta))|h_{t-1})$

How much does the action I take at time t, reduce my entropy over the distribution over $A^*|h_{t-1}$.

Understanding the Information Ratio

Information Ratio

$$\Gamma_t = \frac{\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta) - R(A_t,\theta)]^2}{I(A^*;(A_t, R(A_t,\theta))|h_{t-1})} \tag{9}$$

Suppose the Information ratio Γ_t is bounded by a small constant. What does that mean?

1. Either the algorithm picks an action that will have a small numerator. i.e It will choose the best action given the information it has (**exploit**).

2. Else the algorithm picks an action that will have high denominator. i.e decrease the uncertainty about optimal action A^* (**explore**).

We are typically interested in algorithms with such a small bounded information ratio ($\Gamma_t \leq \bar{\Gamma}$).

A general regret bound for any Online Learning Algorithm

We are interested in the bounding the bayesian regret:

$$\mathbb{E}_{h_{t-1}\sim P_{\mathcal{H}_{t-1}}}\left[\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta)-R(A_t,\theta)]\right] \tag{10}$$

Plugging in the information ratio definition:

$$\begin{aligned}
\text{Bayesian Regret} &= \mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{t=1}^{T}\mathbb{E}_{\theta|h_{t-1}}[R(A^*,\theta)-R(A_t,\theta)]\right] && (11)\\
&= \mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{t=1}^{T}\sqrt{\Gamma_t I(A^*;(A_t,R(A_t,\theta))|h_{t-1})}\right] && (12)\\
\text{(by bounded information ratio)} &\leq \sqrt{\bar{\Gamma}}\,\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{t=1}^{T}\sqrt{I(A^*;(A_t,R(A_t,\theta))|h_{t-1})}\right] && (13)\\
\text{(by Cauchy Schwartz)} &\leq \sqrt{\bar{\Gamma}\,T\,\mathbb{E}_{\mathcal{H}_{t-1}}\left[\sum_{t=1}^{T}I(A^*;(A_t,R(A_t,\theta))|h_{t-1})\right]} && (14)
\end{aligned}$$

A general regret bound for any Online Learning Algorithm

Bayesian Regret is bounded as follows:

$$\text{Bayesian Regret} \leq \sqrt{\bar{\Gamma} T \mathbb{E}_{\mathcal{H}_{t-1}} \left[\sum_{t=1}^{T} I(A^*; (A_t, R(A_t, \theta)) | h_{t-1}) \right]} \tag{15}$$

Recall that $h_{t-1} = (a_1, r_1, a_2, r_2.., a_{t-1}, r_{t-1})$ is a particular history that was observed. We will use $\mathcal{H}_{t-1} = (A_1, R_1, A_2, R_2, .., A_{t-1}, R_{t-1})$ to denote the random variable for history. i.e $\mathcal{H}_{t-1}$ is the possible histories that could have been observed

Let's figure out the inner mutual information term:

$$\mathbb{E}_{\mathcal{H}_{t-1}}[I(A^*; (A_t, R(A_t, \theta)) | h_{t-1})] = I(A^*; (A_t, R(A_t, \theta)) | \mathcal{H}_{t-1}) \tag{16}$$

A general regret bound for any Online Learning Algorithm

Bayesian Regret is bounded as follows:

$$\text{Bayesian Regret} \leq \sqrt{\bar{\Gamma} T \mathbb{E}_{\mathcal{H}_{t-1}} \left[\sum_{t=1}^{T} I(A^*; (A_t, R(A_t, \theta)) | h_{t-1}) \right]} \tag{17}$$

The inner mutual information term is given by:

$$\mathbb{E}_{\mathcal{H}_{t-1}}[I(A^*; (A_t, R(A_t, \theta)) | h_{t-1})] = I(A^*; R(A_t, \theta) | \mathcal{H}_{t-1}) \tag{18}$$

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_{t-1}} \left[\sum_{t=1}^{T} I(A^*; (A_t, Y_t(A_t, \theta)) | h_{t-1}) \right] &= \sum_{t=1}^{T} I(A^*; (A_t, Y_t(A_t, \theta)) | \mathcal{H}_{t-1}) \\
&= \sum_{t=1}^{T} I(A^*; (A_t, R(A_t, \theta)) | (A_1, R_1, .., A_{t-1}, R_{t-1})) \\
\text{(by Mutual Information Chain Rule)} &= I(A^*; (A_1, R_1, A_2, R_2.., A_T, R_T)) \\
&= I(A^*; \mathcal{H}_T)
\end{aligned}$$

A general regret bound for any Online Learning Algorithm

Bayesian Regret is bounded as follows:

$$\text{Bayesian Regret} \leq \sqrt{\bar{\Gamma} T \mathbb{E}_{\mathcal{H}_{t-1}} \left[\sum_{t=1}^{T} I(A^*; (A_t, R(A_t, \theta)) | h_{t-1}) \right]} \tag{19}$$

The inner mutual information term is equal to:

$$\mathbb{E}_{\mathcal{H}_{t-1}} \left[\sum_{t=1}^{T} I(A^*; (A_t, R(A_t, \theta)) | h_{t-1}) \right] = I(A^*; \mathcal{H}_T) \leq H(A^*) \tag{20}$$

Therefore we have the following general bound on bayesian regret for any Online Learning algorithm with bounded information ratio:

General Bound

$$\text{Bayesian Regret} \leq \sqrt{\bar{\Gamma} H(A^*) T} \tag{21}$$

A general regret bound for any Online Algorithm

Lets think what this bound actually says:

General Bound

$$\text{Bayesian Regret} \leq \sqrt{\bar{\Gamma} H(A^*) T} \tag{22}$$

Average mistakes made per iteration:

$$\lim_{T\to\infty} \frac{\text{Bayesian Regret}}{T} \leq \lim_{T\to\infty} \frac{\sqrt{\bar{\Gamma} H(A^*) T}}{T} = \lim_{T\to\infty} \frac{\sqrt{\bar{\Gamma} H(A^*)}}{\sqrt{T}} = 0 \tag{23}$$

If we can find an algorithm that does exploration and exploitation intelligently, i.e has a bounded information ratio ($\bar{\Gamma}$), we have found a algorithm which makes zero mistakes in the long run thereby finding the optimal solution.

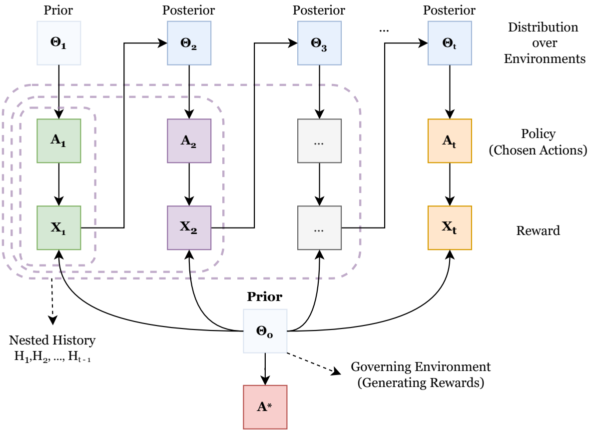
Thompson sampling is precisely one such algorithm which gives zero regret (zero mistakes)!.

Goal: Bounding Information Ratio

$$\Gamma_t = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A_t, \Theta_0)|\mathcal{H}_{t-1}])^2}{I(A^*; (A_t, X_t)|\mathcal{H}_{t-1})} \leq \ ?$$

[1] $X_t = R(A_t, \Theta_0)$

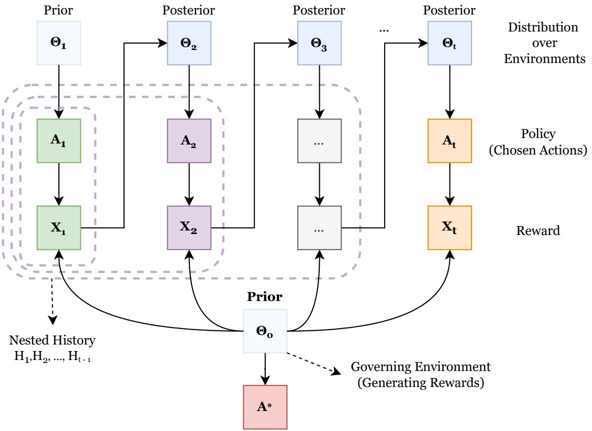
Recap: Information Ratio



$$\Gamma_t = \frac{\left(\mathbb{E}[R(A^*, \Theta_0) - R(A_t, \Theta_0) | \mathcal{H}_{t-1}]\right)^2}{I(A^*; (A_t, X_t) | \mathcal{H}_{t-1})}$$

[1] $X_i = R(A_i, \Theta_0)$

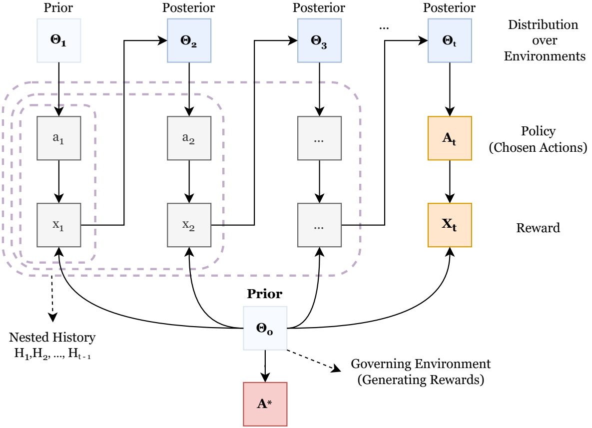
Recap: Information Ratio



$$\Gamma_t(h) = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A_t, \Theta_0) | \mathcal{H}_{t-1} = h])^2}{I(A^*; (A_t, X_t) | \mathcal{H}_{t-1} = h)}$$

[1] $h = (a_1, x_1, \ldots, a_{t-1}, x_{t-1})$

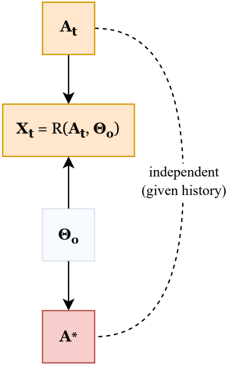
Recap: Information Ratio



$$\Gamma_t(h) = \frac{\left(\mathbb{E}[R(A^*, \Theta_0) - R(A_t, \Theta_0) | \mathcal{H}_{t-1} = h]\right)^2}{I(A^*; (A_t, X_t) | \mathcal{H}_{t-1} = h)}$$

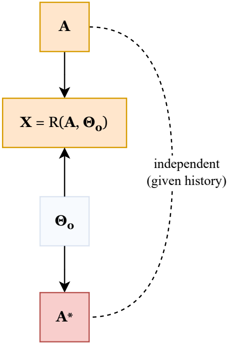
[1] $h = (a_1, x_1, \ldots, a_{t-1}, x_{t-1})$

Information Ratio: Compact Form



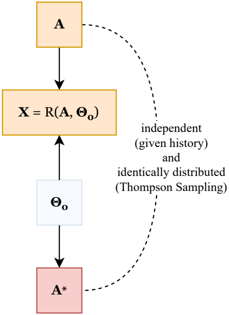
$$\Gamma_t = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A_t, \Theta_0))^2}{I(A^*; (A_t, X_t))}$$

Information Ratio: Compact Form



$$\Gamma = \frac{\left(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)]\right)^2}{I(A^*; (A, X))}$$

Information Ratio: Compact Form



$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

Information Ratio: A Closer Look

$$\Gamma = \frac{\left(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)]\right)^2}{I(A^*; (A, X))}$$

Information Ratio: A Closer Look

$$\Gamma = \frac{\left(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)]\right)^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot \left(\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)]\right)$$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)])$$

$$I(A^*; (A, X)) = \sum_{a \in \mathcal{A}} p_A(a) \cdot \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot D_{\mathrm{KL}}(R(a, \Theta_0) | A^* = a^* \parallel R(a, \Theta_0))$$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)])$$

$$I(A^*; (A, X)) = \sum_{a \in \mathcal{A}} p_A(a) \cdot \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot D_{\mathrm{KL}}(R(a, \Theta_0) | A^* = a^* \parallel R(a, \Theta_0))$$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)])$$

$$I(A^*; (A, X)) = \sum_{a \in \mathcal{A}} p_A(a) \cdot \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot D_{\mathrm{KL}}(R(a, \Theta_0) | A^* = a^* \parallel R(a, \Theta_0))$$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)])$$

$$I(A^*; (A, X)) = \sum_{a \in \mathcal{A}} p_A(a) \cdot \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot D_{\mathrm{KL}}(R(a, \Theta_0) | A^* = a^* \parallel R(a, \Theta_0))$$

Pinsker's Inequality: $(\mathbb{E}[X] - \mathbb{E}[Y])^2 \leq \frac{1}{2} D_{\mathrm{KL}}(X \| Y)$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a] - \mathbb{E}[R(a, \Theta_0)])$$

$$\frac{I(A^*; (A, X))}{2} \geq \sum_{a, a^* \in \mathcal{A}} p_A(a) \cdot p_{A^*}(a^*) \cdot (\mathbb{E}[R(a, \Theta_0) | A^* = a^*] - \mathbb{E}[R(a, \Theta_0)])^2$$

Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (\underbrace{\mathbb{E}[R(a, \Theta_0) | A^* = a]}_{f(a,a)} - \underbrace{\mathbb{E}[R(a, \Theta_0)]}_{g(a)})$$

$$\frac{I(A^*; (A, X))}{2} \geq \sum_{a, a^* \in \mathcal{A}} p_A(a) \cdot p_{A^*}(a^*) \cdot (\underbrace{\mathbb{E}[R(a, \Theta_0) | A^* = a^*]}_{f(a,a^*)} - \underbrace{\mathbb{E}[R(a, \Theta_0)]}_{g(a)})^2$$

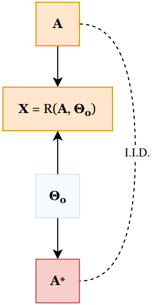
Information Ratio: A Closer Look

$$\Gamma = \frac{(\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)])^2}{I(A^*; (A, X))}$$

$$\mathbb{E}[R(A^*, \Theta_0) - R(A, \Theta_0)] = \sum_{a \in \mathcal{A}} p_A(a) \cdot (f(a, a) - g(a))$$

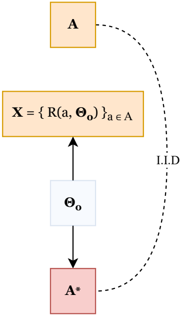
$$\frac{I(A^*; (A, X))}{2} \geq \sum_{a, a^* \in \mathcal{A}} p_A(a) \cdot p_{A^*}(a^*) \cdot (f(a, a^*) - g(a))^2$$

Three Scenarios



General Case

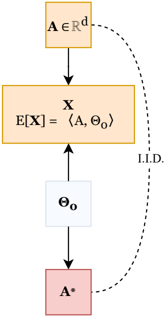
$$\Gamma \leq \frac{|\mathcal{A}|}{2}$$



Full Information

receive all reward distributions regardless of chosen action

$$\Gamma \leq \frac{1}{2}$$



Linear Bandit

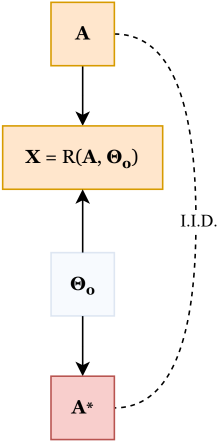
expected reward is a linear function of action

$$\Gamma \leq \frac{d}{2}$$

Scenario 1: General Case

Bound: $\Gamma \leq \dfrac{|\mathcal{A}|}{2}$

Proof.



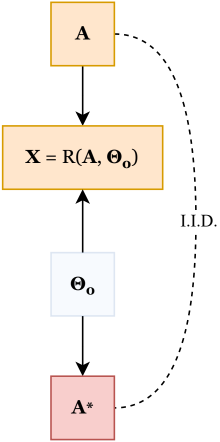
[1] Cauchy-Schwarz: $(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$

Scenario 1: General Case

Bound: $\Gamma \le \frac{|\mathcal{A}|}{2}$

Proof.

$$\text{NUM.} = \left(\sum_{a \in \mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)] \right)^2$$

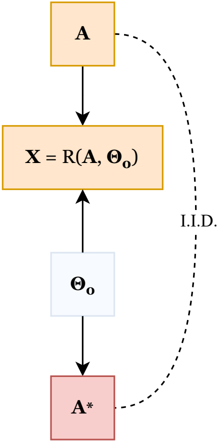


[1] Cauchy-Schwarz: $(x_1 + \cdots + x_n)^2 \le n(x_1^2 + \cdots + x_n^2)$

Bound: $\Gamma \leq \frac{|\mathcal{A}|}{2}$

Proof.

$$\begin{aligned} \text{NUM.} &= \left(\sum_{a \in \mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)] \right)^2 \\ &\overset{(1)}{\leq} |\mathcal{A}| \cdot \sum_{a \in \mathcal{A}} p_A(a)^2 \cdot [f(a,a) - g(a)]^2 \end{aligned}$$

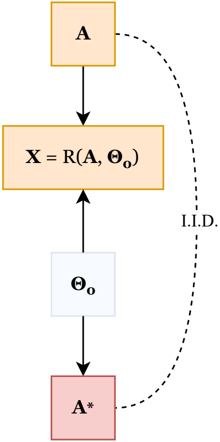


[1] Cauchy-Schwarz: $(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$

Bound: $\Gamma \leq \frac{|\mathcal{A}|}{2}$

Proof.

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a)\cdot[f(a,a)-g(a)]\right)^2$$
$$\overset{(1)}{\leq} |\mathcal{A}|\cdot\sum_{a\in\mathcal{A}} p_A(a)^2\cdot[f(a,a)-g(a)]^2$$
$$\leq |\mathcal{A}|\cdot\sum_{a,a^*\in\mathcal{A}} p_A(a)\cdot p_A(a^*)\cdot[f(a,a^*)-g(a)]^2$$



[1] Cauchy-Schwarz: $(x_1+\cdots+x_n)^2 \leq n(x_1^2+\cdots+x_n^2)$

Bound: $\Gamma \leq \frac{|\mathcal{A}|}{2}$

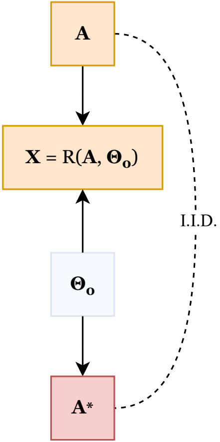
Proof.

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a)\cdot[f(a,a)-g(a)]\right)^2$$

$$\overset{(1)}{\leq} |\mathcal{A}|\cdot\sum_{a\in\mathcal{A}} p_A(a)^2\cdot[f(a,a)-g(a)]^2$$

$$\leq |\mathcal{A}|\cdot\sum_{a,a^*\in\mathcal{A}} p_A(a)\cdot p_A(a^*)\cdot[f(a,a^*)-g(a)]^2$$

$$\leq |\mathcal{A}|\cdot\sum_{a,a^*\in\mathcal{A}} p_A(a)\cdot p_{A^*}(a^*)\cdot[f(a,a^*)-g(a)]^2$$

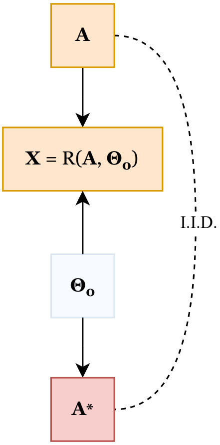


[1] Cauchy-Schwarz: $(x_1+\cdots+x_n)^2 \leq n(x_1^2+\cdots+x_n^2)$

Bound: $\Gamma \le \dfrac{|\mathcal{A}|}{2}$

Proof.

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a)\cdot[f(a,a)-g(a)]\right)^2$$
$$\overset{(1)}{\le} |\mathcal{A}|\cdot\sum_{a\in\mathcal{A}} p_A(a)^2\cdot[f(a,a)-g(a)]^2$$
$$\le |\mathcal{A}|\cdot\sum_{a,a^*\in\mathcal{A}} p_A(a)\cdot p_A(a^*)\cdot[f(a,a^*)-g(a)]^2$$
$$\le |\mathcal{A}|\cdot\sum_{a,a^*\in\mathcal{A}} p_A(a)\cdot p_{A^*}(a^*)\cdot[f(a,a^*)-g(a)]^2$$
$$\le |\mathcal{A}|\cdot\frac{\text{DEN.}}{2}$$

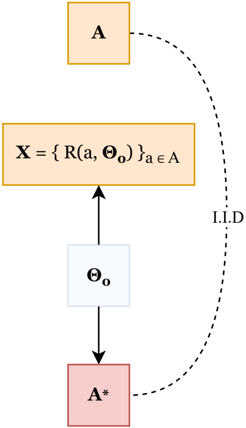


[1] Cauchy-Schwarz: $(x_1+\cdots+x_n)^2\le n(x_1^2+\cdots+x_n^2)$

Scenario 2: Full Information

Bound: $\Gamma \leq \frac{1}{2}$

Proof.



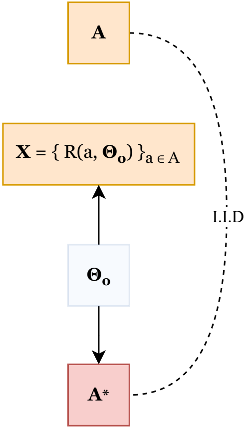
[1]Using Pinsker's inequality

Scenario 2: Full Information

Bound: $\Gamma \leq \dfrac{1}{2}$

Proof.

$$\text{DEN.} = \sum_{a \in \mathcal{A}} I(A^*; R(a, \Theta_0))$$



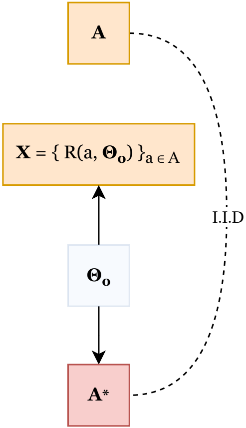
Scenario 2: Full Information

Bound: $\Gamma \leq \frac{1}{2}$

Proof.

$$\text{DEN.} = \sum_{a \in \mathcal{A}} I(A^*; R(a, \Theta_0))$$

$$\overset{(1)}{\geq} 2 \sum_{a \in \mathcal{A}} \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot [f(a, a^*) - g(a)]^2$$



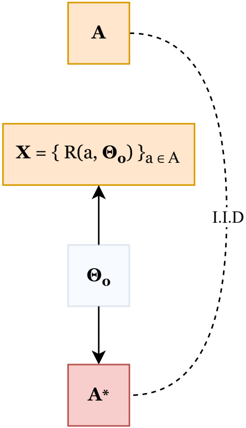
[1] Using Pinsker's inequality

Scenario 2: Full Information

Bound: $\Gamma \leq \frac{1}{2}$

Proof.

$$\text{DEN.} = \sum_{a \in \mathcal{A}} I(A^*; R(a, \Theta_0))$$
$$\overset{(1)}{\geq} 2 \sum_{a \in \mathcal{A}} \sum_{a^* \in \mathcal{A}} p_{A^*}(a^*) \cdot [f(a, a^*) - g(a)]^2$$
$$= 2 \sum_{a, a^* \in \mathcal{A}} p_A(a^*) \cdot [f(a, a^*) - g(a)]^2$$

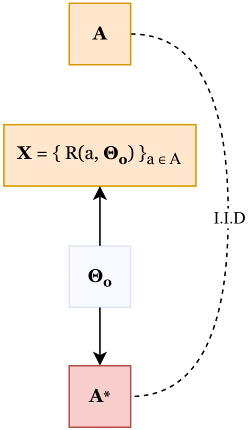


[1] Using Pinsker's inequality

Bound: $\Gamma \leq \frac{1}{2}$

Proof.

$$\text{DEN.} \geq 2 \sum_{(a,a^*)\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$

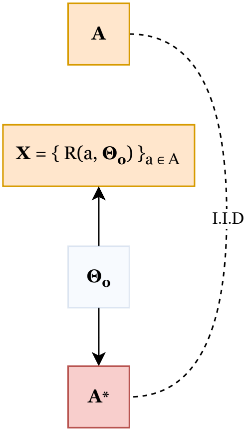


Bound: $\Gamma \leq \dfrac{1}{2}$

Proof.

$$\text{DEN.} \geq 2 \sum_{(a,a^*)\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]\right)^2$$



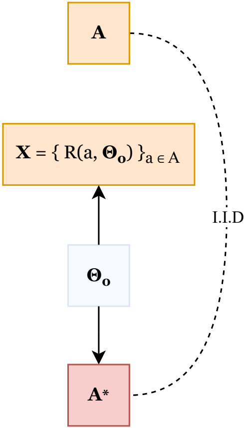
Bound: $\Gamma \leq \frac{1}{2}$

Proof.

$$\text{DEN.} \geq 2 \sum_{(a,a^*)\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]\right)^2$$

$$\overset{(1)}{\leq} \sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]^2$$



[1] Jensen's Inequality: $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$

Scenario 2: Full Information

Bound: $\Gamma \leq \frac{1}{2}$

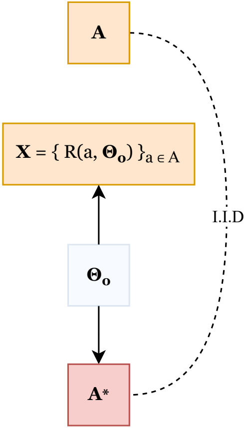
Proof.

$$\text{DEN.} \geq 2 \sum_{(a,a^*)\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$

$$\text{NUM.} = \left(\sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]\right)^2$$

$$\overset{(1)}{\leq} \sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]^2$$

$$\leq \sum_{a,a^*\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$



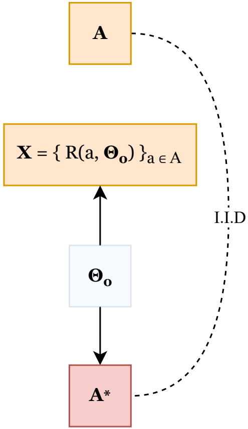
[1] Jensen's Inequality: $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$

Bound: $\Gamma \leq \frac{1}{2}$

Proof.

$$\text{DEN.} \geq 2 \sum_{(a,a^*)\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2$$

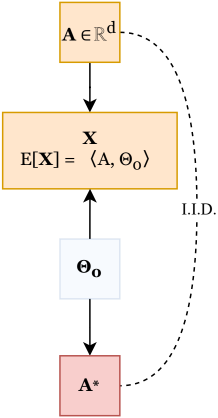
$$\begin{aligned}
\text{NUM.} &= \left(\sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]\right)^2 \\
&\overset{(1)}{\leq} \sum_{a\in\mathcal{A}} p_A(a) \cdot [f(a,a) - g(a)]^2 \\
&\leq \sum_{a,a^*\in\mathcal{A}} p_A(a^*) \cdot [f(a,a^*) - g(a)]^2 \\
&\leq \frac{\text{DEN.}}{2}
\end{aligned}$$



[1] Jensen's Inequality: $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$

Bound: $\Gamma \leq \dfrac{d}{2}$

Proof.

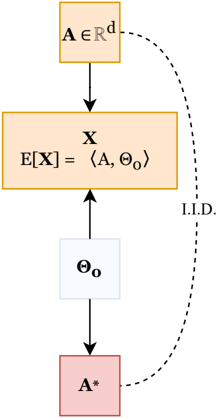


[1] $\text{trace}(X) \leq \sqrt{\text{rank}(X)} \cdot \|X\|_F$

Scenario 3: Linear Bandit with Correlated Arms

Bound: $\Gamma \leq \dfrac{d}{2}$

Proof. Define the matrix $M \in \mathbb{R}^{k \times k}$, where $k = |\mathcal{A}|$ and $M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot (f(a_i, a_j) - g(a_i))$.

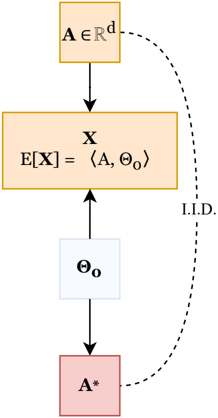


[1] $\text{trace}(X) \leq \sqrt{\text{rank}(X)} \cdot \|X\|_F$

Bound: $\Gamma \leq \dfrac{d}{2}$

Proof. Define the matrix $M \in \mathbb{R}^{k \times k}$, where $k = |\mathcal{A}|$ and $M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot (f(a_i, a_j) - g(a_i))$.

$$\text{NUM.} \overset{(1)}{=} \text{trace}(M)^2$$



[1] $\text{NUM.} = \sum_{a \in \mathcal{A}} p_A(a) \cdot (f(a, a) - g(a))$

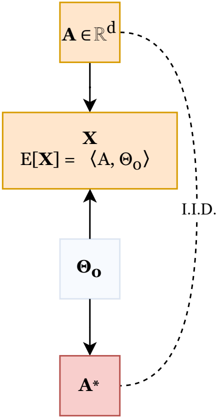
Scenario 3: Linear Bandit with Correlated Arms

Bound: $\Gamma \leq \dfrac{d}{2}$

Proof. Define the matrix $M \in \mathbb{R}^{k \times k}$, where $k = |\mathcal{A}|$ and $M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot (f(a_i, a_j) - g(a_i))$.

$$\text{NUM.} = \text{trace}(M)^2$$

$$\text{DEN.} \stackrel{(1)}{=} 2\|M\|_F^2$$



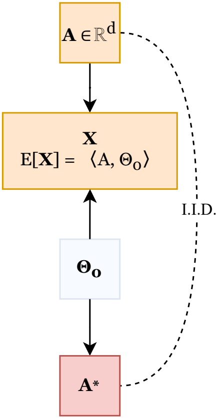
[1] DEN. $\geq 2 \sum_{a, a^* \in \mathcal{A}} p_A(a) \cdot p_{A^*}(a^*) \cdot (f(a, a^*) - g(a))^2$

Scenario 3: Linear Bandit with Correlated Arms

Bound: $\Gamma \leq \dfrac{d}{2}$

Proof. Define the matrix $M \in \mathbb{R}^{k \times k}$, where $k = |\mathcal{A}|$ and $M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot (f(a_i, a_j) - g(a_i))$.

$$\begin{aligned}
\text{NUM.} &= \operatorname{trace}(M)^2 \\
\text{DEN.} &\geq 2\|M\|_F^2 \\
\Gamma &= \frac{\text{NUM.}}{\text{DEN.}} \\
&\leq \frac{\operatorname{trace}(M)^2}{2\|M\|_F^2} \\
&\overset{(1)}{\leq} \frac{\operatorname{rank}(M)}{2}
\end{aligned}$$

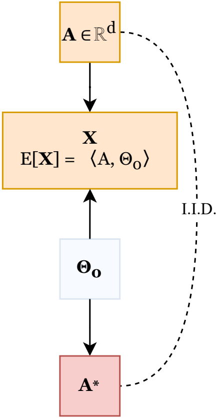


[1] $\operatorname{trace}(X) \leq \sqrt{\operatorname{rank}(X)} \cdot \|X\|_F$

Bound: $\Gamma \leq \frac{d}{2}$

Proof.

$$\Gamma \leq \frac{\text{rank}(M)}{2}$$



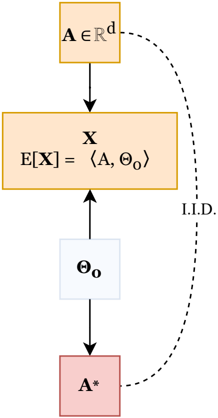
Scenario 3: Linear Bandit

Bound: $\Gamma \leq \frac{d}{2}$

Proof.

$$\Gamma \leq \frac{\text{rank}(M)}{2}$$

$$M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot [\underbrace{f(a_i, a_j)}_{\langle a_i, \theta^j \rangle} - \underbrace{g(a_i)}_{\langle a_i, \theta \rangle}]$$



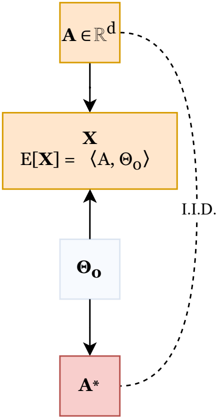
Bound: $\Gamma \leq \frac{d}{2}$

Proof.

$$\Gamma \leq \frac{\text{rank}(M)}{2}$$

$$M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot [\underbrace{f(a_i, a_j)}_{\langle a_i, \theta^j \rangle} - \underbrace{g(a_i)}_{\langle a_i, \theta \rangle}]$$

$$= \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot \langle a_i, \theta^j - \theta \rangle$$



Bound: $\Gamma \leq \frac{d}{2}$

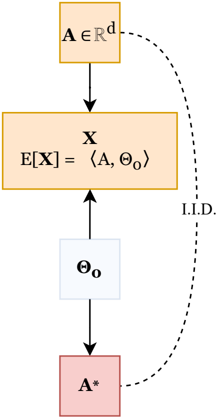
Proof.

$$\Gamma \leq \frac{\text{rank}(M)}{2}$$

$$M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot [\underbrace{f(a_i, a_j)}_{\langle a_i, \theta^j \rangle} - \underbrace{g(a_i)}_{\langle a_i, \theta \rangle}]$$

$$= \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot \langle a_i, \theta^j - \theta \rangle$$

We will show that $\text{rank}(M) \leq d$.



Bound: $\Gamma \leq \frac{d}{2}$

Proof.

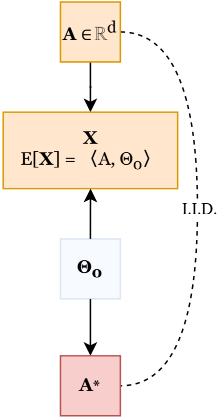
$$\Gamma \leq \frac{\operatorname{rank}(M)}{2}$$

$$M_{ij} := \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot [\underbrace{f(a_i, a_j)}_{\langle a_i, \theta^j \rangle} - \underbrace{g(a_i)}_{\langle a_i, \theta \rangle}]$$

$$= \sqrt{p_A(a_i) \cdot p_A(a_j)} \cdot \langle a_i, \theta^j - \theta \rangle$$

We will show that $\operatorname{rank}(M) \leq d$.

$$M = \underbrace{\begin{bmatrix} \sqrt{p_A(a_1)} \cdot (\theta^1 - \theta)^\top \\ \cdots \\ \sqrt{p_A(a_d)} \cdot (\theta^d - \theta)^\top \end{bmatrix}}_{k \times d} \underbrace{\begin{bmatrix} \sqrt{p_A(a_1)} \cdot a_1 & \ldots & \sqrt{p_A(a_d)} \cdot a_d \end{bmatrix}}_{d \times k}$$



What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
- Full Information: $\Gamma \leq \frac{1}{2}$
- Linear Bandit: $\Gamma \leq \frac{d}{2}$

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
- Linear Bandit: $\Gamma \leq \frac{d}{2}$

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \frac{d}{2}$

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \frac{d}{2}$
  - independent of the number of actions (good!)

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \frac{d}{2}$
  - independent of the number of actions (good!)
  - recall that the regret is bounded by $\sqrt{\Gamma \cdot H(A^*) \cdot T}$

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \frac{d}{2}$
  - independent of the number of actions (good!)
  - recall that the regret is bounded by $\sqrt{\Gamma \cdot H(A^*) \cdot T}$
  - .. when the action space is infinite, $H(A^*)$ can be very large/unbounded!

What Might Go Wrong?

- General Case: $\Gamma \leq \dfrac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \dfrac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \dfrac{d}{2}$
  - independent of the number of actions (good!)
  - recall that the regret is bounded by $\sqrt{\Gamma \cdot H(A^*) \cdot T}$
  - .. when the action space is infinite, $H(A^*)$ can be very large/unbounded!
  - Can we do better?

What Might Go Wrong?

- General Case: $\Gamma \leq \frac{|\mathcal{A}|}{2}$
  - what if the number of actions is very large/infinite?
- Full Information: $\Gamma \leq \frac{1}{2}$
  - ... never happen
- Linear Bandit: $\Gamma \leq \frac{d}{2}$
  - independent of the number of actions (good!)
  - recall that the regret is bounded by $\sqrt{\Gamma \cdot H(A^*) \cdot T}$
  - .. when the action space is infinite, $H(A^*)$ can be very large/unbounded!
  - Can we do better? Yes!

Recap: Linear Stochastic Bandit

In a linear stochastic bandit:

- Environment is parameterized by $\theta \in \Theta \subset \mathbb{R}^d$ with prior P
- Player can choose action $a \in \mathcal{A} \subset \mathbb{R}^d$
- Reward $R(a, \theta) \in [0, 1]$ with $\mathbb{E}[R(a, \theta)] = \langle a, \theta \rangle$

Recap: Linear Stochastic Bandit

In a linear stochastic bandit:

- Environment is parameterized by $\theta \in \Theta \subset \mathbb{R}^d$ with prior P
- Player can choose action $a \in \mathcal{A} \subset \mathbb{R}^d$
- Reward $R(a, \theta) \in [0, 1]$ with $\mathbb{E}[R(a, \theta)] = \langle a, \theta \rangle$

The Nature samples $\theta^* \sim P$. Then for $t = 1, \ldots, T$:

1. Sample θ_t from the posterior $\theta^* \mid \mathcal{H}_{t-1}$
2. Choose $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle$
3. Receive reward $R(A_t, \theta^*)$

In a linear stochastic bandit:

- Environment is parameterized by $\theta \in \Theta \subset \mathbb{R}^d$ with prior P
- Player can choose action $a \in \mathcal{A} \subset \mathbb{R}^d$
- Reward $R(a, \theta) \in [0, 1]$ with $\mathbb{E}[R(a, \theta)] = \langle a, \theta \rangle$

The Nature samples $\theta^* \sim P$. Then for $t = 1, \ldots, T$:

1. Sample θ_t from the posterior $\theta^* \mid \mathcal{H}_{t-1}$
2. Choose $A_t = \text{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle$
3. Receive reward $R(A_t, \theta^*)$

The Bayesian regret of TS is given by

$$\mathrm{BR}_T = \mathop{\mathbb{E}}_{\theta^* \sim P}\left[\sum_{t=1}^{T}\Big(\underbrace{R(A^*, \theta^*)}_{\text{optimal reward}} - \underbrace{R(A_t, \theta^*)}_{\text{player's reward}}\Big)\right]$$

where $A^* = \text{argmax}_{a \in \mathcal{A}} \langle a, \theta^* \rangle$

The Curse of Many Actions

- We have seen that TS achieves

$$\mathrm{BR}_T \le \sqrt{H(A^*)\bar{\Gamma} T} \le \sqrt{\frac{H(A^*)dT}{2}} \le \sqrt{\frac{\log|\mathcal{A}|dT}{2}}$$

- However, this regret bound explodes when $|\mathcal{A}| \to \infty$

The Curse of Many Actions

- We have seen that TS achieves

$$\mathrm{BR}_T \leq \sqrt{H(A^*)\bar{\Gamma}T} \leq \sqrt{\frac{H(A^*)dT}{2}} \leq \sqrt{\frac{\log|\mathcal{A}|dT}{2}}$$

- However, this regret bound explodes when $|\mathcal{A}| \to \infty$
- In the following, we will show that

$$\mathrm{BR}_T = \mathcal{O}(d\sqrt{T\log T}),$$

which is independent of $|\mathcal{A}|$!

The Curse of Many Actions

- We have seen that TS achieves

$$\mathrm{BR}_T \le \sqrt{H(A^*)\bar{\Gamma} T} \le \sqrt{\frac{H(A^*) d T}{2}} \le \sqrt{\frac{\log |\mathcal{A}| d T}{2}}$$

- However, this regret bound explodes when $|\mathcal{A}| \to \infty$
- In the following, we will show that

$$\mathrm{BR}_T = \mathcal{O}(d\sqrt{T \log T}),$$

which is independent of $|\mathcal{A}|$!

- The analysis is based on:

Shi Dong and Benjamin Van Roy. An Information-Theoretic Analysis for Thompson Sampling with Many Actions, *NeurIPS* 2018.

Learning a Near-Optimal Action

- Two key points in the information-theoretic analysis

Learning a Near-Optimal Action

- Two key points in the information-theoretic analysis
  - The information ratio is bounded:

$$\Gamma_t \triangleq \frac{\mathbb{E}_{t-1}[(R(A^*, \theta^*) - R(A_t, \theta^*))^2]}{I_{t-1}(A^*; (A_t, R_t))} \leq \frac{d}{2}$$

Learning a Near-Optimal Action

- Two key points in the information-theoretic analysis
  - The information ratio is bounded:

$$\Gamma_t \triangleq \frac{\mathbb{E}_{t-1}[(R(A^*, \theta^*) - R(A_t, \theta^*))^2]}{I_{t-1}(A^*; (A_t, R_t))} \leq \frac{d}{2}$$

  - The accumulated information gain is bounded:

$$\sum_{t=1}^{T} \mathbb{E}[I_{t-1}(A^*; (A_t, R_t))] = I(A^*; \mathcal{H}_T) \leq H(A^*)$$

Learning a Near-Optimal Action

- Two key points in the information-theoretic analysis
  - The information ratio is bounded:

$$\Gamma_t \triangleq \frac{\mathbb{E}_{t-1}[(R(A^*, \theta^*) - R(A_t, \theta^*))^2]}{I_{t-1}(A^*; (A_t, R_t))} \leq \frac{d}{2}$$

  - The accumulated information gain is bounded:

$$\sum_{t=1}^{T} \mathbb{E}[I_{t-1}(A^*; (A_t, R_t))] = I(A^*; \mathcal{H}_T) \leq H(A^*)$$

- The problem: learning the exact optimal action A^* requires a lot information

Learning a Near-Optimal Action

- Two key points in the information-theoretic analysis
  - The information ratio is bounded:

$$\Gamma_t \triangleq \frac{\mathbb{E}_{t-1}[(R(A^*,\theta^*) - R(A_t,\theta^*))^2]}{I_{t-1}(A^*;(A_t,R_t))} \leq \frac{d}{2}$$

  - The accumulated information gain is bounded:

$$\sum_{t=1}^{T} \mathbb{E}[I_{t-1}(A^*;(A_t,R_t))] = I(A^*;\mathcal{H}_T) \leq H(A^*)$$

- The problem: learning the exact optimal action A^* requires a lot information
- Key idea: maybe we can settle for a near-optimal action?

$$\tilde{\Gamma}_t \triangleq \frac{\mathbb{E}_{t-1}[(R(\tilde{A}_t^*,\theta^*) - R(A_t,\theta^*))^2]}{I_{t-1}(\tilde{A}_t^*;(A_t,R_t))}$$

Quantization of Action Space

- Assume that $\max_{a \in \mathcal{A}} \|a\|_2 \leq 1$ and $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$
- We can find a partition $\mathcal{A} = \cup_{k=1}^{K} \mathcal{A}_k$ such that

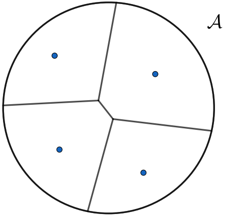
$$\max_{a, a' \in \mathcal{A}_k} \|a - a'\|_2 \leq \epsilon, \quad \forall k = 1, \ldots, K,$$

and

$$K \leq \left(1 + \frac{4}{\epsilon}\right)^d$$

- Define ψ as the index of the partition containing A^*

$$\psi = k \quad \Leftrightarrow \quad A^* \in \mathcal{A}_k$$



Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$
  - $\tilde{A}_t^*$ and A^* share the same distribution

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$
  - $\tilde{A}_t^*$ and A^* share the same distribution $\Rightarrow$ so are $\tilde{A}_t^*$ and A_t

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$
  - $\tilde{A}_t^*$ and A^* share the same distribution $\Rightarrow$ so are $\tilde{A}_t^*$ and A_t
  - Given $\mathcal{H}_{t-1}$ and ψ, $\tilde{A}_t^* \perp A^*, \theta^*, A_t, R_t$

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$
  - $\tilde{A}_t^*$ and A^* share the same distribution $\Rightarrow$ so are $\tilde{A}_t^*$ and A_t
  - Given $\mathcal{H}_{t-1}$ and ψ, $\tilde{A}_t^* \perp A^*, \theta^*, A_t, R_t \Rightarrow (A_t, R_t) \rightarrow \psi \rightarrow \tilde{A}_t^*$

Quantization of Action Space (contd.)

- We define a "blurred" optimal action $\tilde{A}_t^*$ that
  - $\tilde{A}_t^*$ and A^* belong to the same partition;
  - $\tilde{A}_t^*$ and A^* are i.i.d. within the partition
- More precisely:

$$\mathbb{P}_{t-1}(\tilde{A}_t^* = a' \mid A^* = a) = \begin{cases} \mathbb{P}_{t-1}(A^* = a' \mid \psi = k) & \text{if } a, a' \in \mathcal{A}_k \\ 0 & \text{otherwise} \end{cases}$$

- Properties:
  - $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon$
  - $\tilde{A}_t^*$ and A^* share the same distribution $\Rightarrow$ so are $\tilde{A}_t^*$ and A_t
  - Given $\mathcal{H}_{t-1}$ and ψ, $\tilde{A}_t^* \perp A^*, \theta^*, A_t, R_t \;\Rightarrow\; (A_t, R_t) \rightarrow \psi \rightarrow \tilde{A}_t^*$

$$I(\tilde{A}_t^*; (A_t, R_t) \mid \mathcal{H}_{t-1}) \leq I(\psi; (A_t, R_t) \mid \mathcal{H}_{t-1})$$

Regret Bound

Implications:

- $\|\tilde{A}_t^* - A^*\|_2 \le \epsilon \quad \Rightarrow \quad \tilde{A}_t^*$ is near-optimal

$$\begin{aligned} &\mathbb{E}_{t-1}[R(A^*, \theta^*) - R(\tilde{A}_t^*, \theta^*)] \\ &= \mathbb{E}_{t-1}[\langle A^* - \tilde{A}_t^*, \theta^* \rangle] \le \mathbb{E}_{t-1}[\|A^* - \tilde{A}_t^*\| \|\theta^*\|] \le \epsilon \end{aligned}$$

Regret Bound

Implications:

- $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon \quad \Rightarrow \quad \tilde{A}_t^*$ is near-optimal

$$\begin{aligned}&\mathbb{E}_{t-1}[R(A^*, \theta^*) - R(\tilde{A}_t^*, \theta^*)]\\ &= \mathbb{E}_{t-1}[\langle A^* - \tilde{A}_t^*, \theta^* \rangle] \leq \mathbb{E}_{t-1}[\|A^* - \tilde{A}_t^*\|\|\theta^*\|] \leq \epsilon\end{aligned}$$

- $\tilde{A}_t^*$ and A_t share the same distribution $\Rightarrow$ $\tilde{\Gamma}_t$ can be bounded as Γ_t

$$\tilde{\Gamma}_t \triangleq \frac{\mathbb{E}_{t-1}[(R(\tilde{A}_t^*, \theta^*) - R(A_t, \theta^*))^2]}{I_{t-1}(\tilde{A}_t^*; (A_t, R_t))} \leq \frac{d}{2}$$

Regret Bound

Implications:

- $\|\tilde{A}_t^* - A^*\|_2 \leq \epsilon \quad \Rightarrow \quad \tilde{A}_t^*$ is near-optimal

$$\begin{aligned}&\mathbb{E}_{t-1}[R(A^*,\theta^*) - R(\tilde{A}_t^*,\theta^*)] \\ &= \mathbb{E}_{t-1}[\langle A^* - \tilde{A}_t^*, \theta^*\rangle] \leq \mathbb{E}_{t-1}[\|A^* - \tilde{A}_t^*\|\|\theta^*\|] \leq \epsilon\end{aligned}$$

- $\tilde{A}_t^*$ and A_t share the same distribution $\Rightarrow$ $\tilde{\Gamma}_t$ can be bounded as Γ_t

$$\tilde{\Gamma}_t \triangleq \frac{\mathbb{E}_{t-1}[(R(\tilde{A}_t^*,\theta^*) - R(A_t,\theta^*))^2]}{I_{t-1}(\tilde{A}_t^*;(A_t,R_t))} \leq \frac{d}{2}$$

- $I(\tilde{A}_t^*;(A_t,R_t) \mid \mathcal{H}_{t-1}) \leq I(\psi;(A_t,R_t) \mid \mathcal{H}_{t-1}) \;\Rightarrow$
The accumulated information gain is bounded by $H(\psi)$

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[I_{t-1}(\tilde{A}_t^*;(A_t,R_t))] &= \sum_{t=1}^{T} I(\tilde{A}_t^*;(A_t,R_t) \mid \mathcal{H}_{t-1}) \\ &\leq \sum_{t=1}^{T} I(\psi;(A_t,R_t) \mid \mathcal{H}_{t-1}) = I(\psi \mid \mathcal{H}_T) \leq H(\psi)\end{aligned}$$

Regret Bound (contd.)

Putting all pieces toghther, we have

$$\mathrm{BR}_T \leq \underbrace{\sqrt{\frac{H(\psi)dT}{2}}}_{H(\psi) \text{ replaces } H(A^*)} + \underbrace{\epsilon T}_{\text{quantization error}}$$

Regret Bound (contd.)

Putting all pieces toghther, we have

$$\mathrm{BR}_T \leq \underbrace{\sqrt{\frac{H(\psi)dT}{2}}}_{H(\psi) \text{ replaces } H(A^*)} + \underbrace{\epsilon T}_{\text{quantization error}}$$

$$\leq \sqrt{\frac{\log(K)dT}{2}} + \epsilon T \qquad \Leftarrow \text{since } K \leq \left(1 + \frac{4}{\epsilon}\right)^d$$

Putting all pieces toghther, we have

$$\mathrm{BR}_T \leq \underbrace{\sqrt{\frac{H(\psi)dT}{2}}}_{\color{red}{H(\psi)\text{ replaces }H(A^*)}} + \underbrace{\epsilon T}_{\color{red}{\text{quantization error}}}$$

$$\leq \sqrt{\frac{\log(K)dT}{2}} + \epsilon T \qquad \Leftarrow \text{since } K \leq \left(1 + \frac{4}{\epsilon}\right)^d$$

$$\leq d\sqrt{\frac{T}{2}\log\left(1 + \frac{4}{\epsilon}\right)} + \epsilon T$$

Regret Bound (contd.)

Putting all pieces toghther, we have

$$\begin{aligned}\mathrm{BR}_T &\le \underbrace{\sqrt{\frac{H(\psi)dT}{2}}}_{H(\psi)\text{ replaces }H(A^*)} + \underbrace{\epsilon T}_{\text{quantization error}} \\ &\le \sqrt{\frac{\log(K)dT}{2}} + \epsilon T && \Leftarrow \text{since } K \le \left(1+\frac{4}{\epsilon}\right)^d \\ &\le d\sqrt{\frac{T}{2}\log\left(1+\frac{4}{\epsilon}\right)} + \epsilon T\end{aligned}$$

By taking $\epsilon = d/\sqrt{2T}$, we obtain

$$\mathrm{BR}_T \le d\sqrt{\frac{T}{2}}\left(\sqrt{\log\left(1+\frac{4\sqrt{2T}}{d}\right)}+1\right) = \mathcal{O}(d\sqrt{T\log T})$$

Link with Rate Distortion Theory

- A classical problem: how to quantize a random variable X
  - Distortion: $d(X, \tilde{X})$ by some distortion metric
  - Rate: $H(\tilde{X})$ bits to represent $\tilde{X}$

Link with Rate Distortion Theory

- A classical problem: how to quantize a random variable X
  - Distortion: $d(X, \tilde{X})$ by some distortion metric
  - Rate: $H(\tilde{X})$ bits to represent $\tilde{X}$
- In the analysis of TS: how to approximate the optimal action A^*
  - Distortion: the suboptimality $\mathbb{E}[R(A^*, \theta^*) - R(\tilde{A}^*, \theta^*)]$
  - Rate: $H(\psi)$ bits to learn ψ

Link with Rate Distortion Theory

- A classical problem: how to quantize a random variable X
  - Distortion: $d(X, \tilde{X})$ by some distortion metric
  - Rate: $H(\tilde{X})$ bits to represent $\tilde{X}$
- In the analysis of TS: how to approximate the optimal action A^*
  - Distortion: the suboptimality $\mathbb{E}[R(A^*, \theta^*) - R(\tilde{A}^*, \theta^*)]$
  - Rate: $H(\psi)$ bits to learn ψ
- By optimizing over the quantization scheme, we can actually show that

$$\mathrm{BR}_T \leq \underbrace{\sqrt{\frac{R(\epsilon) dT}{2}}}_{R(\epsilon) \text{ replaces } H(A^*)} + \underbrace{\epsilon T}_{\text{quantization error}}$$

Link with Rate Distortion Theory

- A classical problem: how to quantize a random variable X
  - Distortion: $d(X, \tilde{X})$ by some distortion metric
  - Rate: $H(\tilde{X})$ bits to represent $\tilde{X}$
- In the analysis of TS: how to approximate the optimal action A^*
  - Distortion: the suboptimality $\mathbb{E}[R(A^*, \theta^*) - R(\tilde{A}^*, \theta^*)]$
  - Rate: $H(\psi)$ bits to learn ψ
- By optimizing over the quantization scheme, we can actually show that

$$\mathrm{BR}_T \leq \underbrace{\sqrt{\frac{R(\epsilon)dT}{2}}}_{R(\epsilon) \text{ replaces } H(A^*)} + \underbrace{\epsilon T}_{\text{quantization error}}$$

- Remark: the TS algorithm is unchanged; $\tilde{A}$ is purely theoretical

Summary

- A crash course on bandit learning

$$\text{BayesianReg}_T = \mathop{\mathbb{E}}_{\theta^* \sim P}\left[\text{Reg}_T(\pi, \theta^*) = \sum_{t=1}^{T}\Big(\underbrace{R(A^*, \theta^*)}_{\text{optimal reward}} - \underbrace{R(A_t, \theta^*)}_{\text{player's reward}} \Big)\right]$$

- Thompson sampling: playing action via probability matching

$$\mathbb{P}(A_t = \cdot \mid \mathcal{H}_{t-1}) = \mathbb{P}(A^* = \cdot \mid \mathcal{H}_{t-1})$$

- Information ratio: immediate reward v.s. information gain

$$\Gamma_t \triangleq \frac{\mathbb{E}_{t-1}[(R(A^*, \theta^*) - R(A_t, \theta^*))^2]}{I_{t-1}(A^*; (A_t, R_t))} \leq \bar{\Gamma} \quad \Rightarrow \quad \text{BR}_T \leq \sqrt{\bar{\Gamma} H(A^*) T}$$

- Quantization comes to rescue when the action space is large